Hearing Research 470 (2026) 109490

Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Research paper

Exploring the relationship between task difficulty, head-related transfer function and spatial release from masking in a speech-on-speech experiment

Thibault Vicente [a,*], Daniel González-Toledo [b], María Cuevas-Rodríguez [b], Luis Molina-Tanco [b], Arcadio Reyes-Lecuona [b], Lorenzo Picinali [a]

[a] Dyson School of Engineering, Imperial College London, South Kensington Campus, London, SW7 2AZ, United Kingdom
[b] Telecommunication Research Institute (TELMA), Universidad de Málaga, Málaga, 29071, Spain

ARTICLE INFO

Keywords:
Spatial audio
Intelligibility
Task difficulty

ABSTRACT

It is known that individuals make use of spatial hearing cues to improve the audibility of a target signal and separate it from competing sounds. This phenomenon is known as spatial release from masking (SRM). Recent research has shown that this happens also when sources are located in the median plane, where interaural differences are limited. When assessing this within virtual conditions, it has been shown that employing individually measured head-related transfer functions (HRTFs) results in higher SRM abilities compared to using non-individual filters. In a previously published work, we found that Spanish speakers benefit from individual HRTFs when discriminating a target English speech from a single masker in the median plane. This study replicates the protocol of that previous work, varying the number of maskers and participants' English proficiency levels to explore relationships among task difficulty and HRTF use. Results from a first experiment show that English speakers behave differently to Spanish ones; their SRM advantage is not significant. We suggest that this is due to their language proficiency, which allows them to rely on spectral glimpsing alone, that is, exploiting spectro-temporal gaps between voices rather than spectral cues introduced by spatial separation. A second experiment introduces a second speech masker, co-located with the first; by making the task more complex, participants seem to increase their reliance on spatial cues, resulting in significant effects of masker position and HRTF. This highlights a trade-off between the use of target glimpsing and spatial cues and the need for further exploration into how task difficulty influences SRM with different HRTFs.

1. Introduction

Understanding speech in a noisy background is a common yet challenging daily task, often referred to as "cocktail-party" (Cherry, 1953). In these contexts, listeners must isolate the target speech from the competing sources to focus their attention effectively (e.g., Bronkhorst, 2000, 2015). The success of this auditory stream segregation depends on the degree of masking caused by competing sources on target. This masking can be broken down into a combination of energetic masking (EM) and informational masking (IM). On one hand, EM occurs because of spectro-temporal overlaps of the target and masker signals at the peripheral level. These overlaps decrease target audibility in specific spectro-temporal regions, which degrades the segregation of the target from the maskers thus reducing intelligibility. On the other hand, IM arises from perceptual similarities between target and maskers that introduce target uncertainties preventing the listener from determining *what, when, where,* or *who* to attend to even when audibility is guaranteed. Differences between the target and masker signals generally enhance the release from IM (e.g., Brungart, 2001), as target segregation becomes easier under such conditions. These differences cause spectro-temporal regions wherein target information is less affected by maskers, so-called target glimpsing. However, EM is also affected by altering the spectro-temporal overlap between the target and masker, thereby intertwining the effects of the two types of release from masking.

Speech recognition is affected by linguistic factors such as the listener's language background, especially when the target is in the midst of other competing sound sources. Listening to a non-native language (language learned some years after acquired your first language) is challenging and generally intelligibility is degraded in cocktail-party situations. Especially, competing speech seems to increase the amount of IM for Non-natives more than for Natives (see reviews from Garcia Lecumberri et al., 2010; Scharenborg and van Os, 2019). This amount of masking can be reduced by the listener's language proficiency and/or the age at which the language was learned.

* Corresponding author.
E-mail address: thibault.vicente@univ-lemans.fr (T. Vicente).

https://doi.org/10.1016/j.heares.2025.109490
Received 9 September 2025; Received in revised form 8 November 2025; Accepted 25 November 2025
Available online 6 December 2025

Listener's spatial hearing abilities are also essential for intelligibility. Target and maskers being located in different positions can significantly benefit to the listeners. This benefit, known as spatial release from masking (SRM), is thought to be based mainly on the use of interaural differences, both spectral and temporal. Although monaural benefits to SRM in horizontal-plane listening have been well established (e.g., Freyman et al., 1999, 2001), recent work has demonstrated that monaural cues can also play a significant role in SRM when sources are presented in the median plane (Martin et al., 2012; González-Toledo et al., 2024).

When sources are positioned in the horizontal plane, the signals arrive at the listener's ears with interaural time differences (ITD) and interaural level differences (ILD). Interaural differences are minimal (but not null due to slight body asymmetry) when sources are located at 0° or 180° azimuth. Differences in ILD between target speech and masking sources lead to different signal-to-noise ratios (SNRs) at the two ears, and the one with the better SNR therefore provides more information related to the target. This is commonly called better-ear listening. Moreover, differences in ITD between target and masking sources are also relevant, thanks to the hearing system's ability to use phase differences to cancel the masking signals and internally improve the SNR. This was explained by Durlach (1972) in his Equalization-Cancellation theory. It should be noted, however, that this theory represents only one possible explanation of how spatial cues enhance speech-in-speech recognition. Other works (e.g., Freyman et al., 2001; Kidd et al., 2016) have shown that perceived spatial separation provides relevant cues for effective auditory stream segregation, facilitating selective attention to the target and inhibition of the masker. In this view, spatial differences contribute to both IM and EM release.

When sources are positioned in the median plane (i.e., the symmetry plane perpendicular to the interaural axis, splitting the human body into two symmetric halves), the differences in ILD and ITD between target and masker are minimal—yet existing due to slight head asymmetry, while spectral differences can be observed in terms of spatial cues. Both monaural and interaural cues rely on anthropomorphic features of the human body. For instance, monaural cues are known to be influenced by the ear shape, while interaural cues by the distance between the entrance of the ear canals (i.e., the size of the head). These cues can be captured via a filtering function called head-related transfer function (HRTF), which is effectively the transfer function from one spatial location to the listener's ear. HRTFs are useful for binaural reproduction in simulation of realistic immersive soundfields via a pair of headphones. Based on morphological features, HRTFs are specific to each individual (Picinali and Katz, 2023). Using individually-measured HRTFs allows to increase sound source rendering, thus, the localisation accuracy these sources (e.g., Jenny and Reuter, 2020).

Assessing speech intelligibility with various HRTFs leads to differences in SRM in the horizontal plane, as binaural cues vary from one HRTF to the other (Cuevas-Rodríguez et al., 2019). What happens when sources are separated only in the median plane is a whole different matter. In the recent years, a few studies have been carried out further exploring the effect of SRM in the median plane. For example, Berwick and Lee (2020) observed SRM using noise maskers, when sources were located in the median plane. Using a loudspeaker array, they found a significant interaction between masker location and speech reception threshold (SRT), more specifically an advantage of up to 3.5 dB with respect to the co-located condition (i.e. target and masker in the same position). Exploring a similar effect in the binaural audio domain (i.e. using headphones rather than loudspeakers), McAnally et al. (2002) replicated a previous study by Bolia et al. (1999) using individual HRTFs, with residual ITD explicitly removed, in a speech-on-speech task. An SRM of about 1.3 dB was measured when the target and masker were at different elevations in the median plane (two different same-sex speakers were used as stimuli), demonstrating the relevance of spectral cues for SRM in the median plane. In a later study, Martin et al. (2012) repeated this experiment, further investigating the effect of HRTF asymmetries on SRM in the median plane. In addition to the individual HRTF, they employed three other HRTFs: one with two left ears, one with two right ears, and another with averages of both ears. The individual HRTF led to higher SRM compared to any of the manipulated HRTFs.

Monaural spatial cues are crucial for localisation in the median plane, and they are highly dependent on the individual morphology of ears, head and torso. This suggests that, when individual HRTFs are degraded, or non-individual HRTFs are used, the SRM is significantly lower in the median plane. These conditions were explicitly assessed in a previous study from the authors of this manuscript (González-Toledo et al., 2024). The protocol proposed by Martin et al. (2012) was extended to investigate the effect of employing individual HRTF on SRM in the median plane. The target and masker were taken from the Coordinate Measure Response (CRM, Bolia et al., 2000), and played simultaneously. The sources were spatialised through headphones at −50°, 0° or +50° of elevation, in front of the listener. While in the co-located condition the percentage of correct words was comparable between individual and non-individual HRTFs, the former outperformed the latter when target and masker were spatially separated. More specifically, in the individual HRTF condition the separation resulted in an improvement of 12.5% of the SRM, while in the non-individual condition, this improvement was limited to 6.5%. The data were gathered by recruiting only native Spanish speakers, but using the English corpus, which may have influenced the amount of IM in the experiment.

The results from González-Toledo et al. (2024) and Martin et al. (2012) suggest that unprocessed individual HRTFs provide larger SRM in the median plane. This is observed because IM or EM decreased in the spatially separated conditions (as opposed to another HRTF). Considering unprocessed individual HRTFs allow a better binaural reproduction realism, thereby the target uncertainty decreases, leading to a lower IM. The amount of EM likely varies between two HRTF conditions because the monaural spatial cues are HRTF-dependent. Then, an HRTF providing more cues should provide higher SRM. Furthermore, González-Toledo et al. (2024) recruited only Non-native English speakers (Non-native group), which may have introduced some additional IM. Then this effect may have interacted with the HRTF individualisation.

In recent years, there has been growing interest in finding alternatives to acoustically measuring HRTFs for personalised binaural rendering. Attempts have been made to rely on behavioural experiments in order to select a best-fitting non-individual HRTF for a given subject, but such methods have been shown to be often problematic (e.g., Kim et al., 2020). Recent approaches successfully employed numerical methods, including computational models mimicking spatial hearing perception (e.g. Daugintis et al., 2023).

The study aims to investigate the relationship between task difficulty — manipulated by varying the number of maskers (Rosen et al., 2013; Freyman et al., 2004), and the native language of the participants (Borghini and Hazan, 2018; Garcia Lecumberri et al., 2010; Scharenborg and van Os, 2019) — the use of individual monaural spatial cues, and the reliance on target glimpsing (i.e., spectro-temporal differences between target and masker voices). In order to quantify the difference between HRTFs in a consistent and repeatable manner as previous studies (e.g. Daugintis et al., 2023; Cuevas-Rodríguez et al., 2019), the present study employs two main metrics, which have already been employed in González-Toledo et al. (2024): spectral distortion (SD) and the output of a SRM model proposed by Jelfs et al. (2011). Thereby, the relevance of these metrics to assess differences in EM between HRTFs is investigated. In addition, a glimsping model, inspired by Wasiuk et al. (2022, 2023), is integrated into the analysis to compute spectro-temporal differences between target and masker voices. Various studies showed that speech-on-speech intelligibility can be predicted using a glimpsing model that assesses the proportion of audible glimpses (Wasiuk et al., 2023, 2022; Best et al., 2017).

A first experiment was driven by the interest in exploring the same protocol and conditions of González-Toledo et al. (2024), but this time recruiting native English speakers (Native group). The outcome (see 3.1) was not in line with what was reported in González-Toledo et al. (2024), neither with Martin et al. (2012), the main dissimilarity being that no significant differences could be found between individual and non-individual HRTFs in spatially separated conditions. This prompted the design of a second experiment to further investigate this discrepancy. The second experiment revealed important findings about the use of spatial cues by individuals when faced with tasks of varying difficulty.

2. General methods and materials

2.1. Stimuli

The CRM corpus described by Bolia et al. (2000) was used for both experiments. It consists of sentences following the same structure: "Ready *call sign*, go to *colour number* now". There are eight call signs, four colours and eight numbers resulting in 256 sentences when all combinations are considered. The corpus is recorded with 4 male speakers and 4 female speakers. As in Martin et al. (2012), the original CRM recordings were used. In these recordings, sentences were band-pass filtered from 200 Hz to 18 kHz (the publicly available version being low-pass filtered at 8 kHz) and adjusted to have the same root-mean-square amplitude.

Both target and masker sentences were taken from the CRM corpus. Sentences were presented at a sound level of approximately 60 dBA, and the target-to-masker ratio was 0 dB. Sound source spatialisation was designed to approximate anechoic conditions, using either individual or KEMAR HRTFs. The room was large and acoustically dry, which allowed effective windowing of the HRIRs to suppress reflections and closely simulate anechoic conditions. Details of this process are reported in González-Toledo et al. (2024) for the first experiment and Engel et al. (2023) for the second. In addition, the headphone transfer function was measured for each individual involved in either of the two experiments to derive a headphone equalisation filter, as described by Engel et al. (2022), which was then used for headphone compensation.

2.2. Procedure

The participants visited the premises of Imperial College London twice for Experiment 1. Their HRTFs were measured during a first appointment, and they performed the test during a second appointment. The appointments could be scheduled consecutively. For Experiment 2, both the HRTF measurement and the speech intelligibility test were performed during a single session. Participants were informed about the purpose of the experiment when they first met the experimenter. All provided their written consent. Participation was voluntary and no monetary compensation was provided.

The procedure was similar to the one used by González-Toledo et al. (2024), and summarised here for convenience. An application was developed specifically for the series of experiments, using the 3DTI Toolkit (Cuevas-Rodríguez et al., 2019), a C++ open-source library for real-time binaural spatialisation. During the experiment, participants were seated, with headphones on, in the same room where the HRIRs were measured, in front of a monitor, and asked to interact with a user interface using a mouse to select their answers. The procedure was automatically sequenced for the entire experiment, without intervention of the experimenter. A MOTU UltraLite-mk3 hybrid audio interface was used to play back the stimuli, which were presented binaurally via a pair of Sennheiser HD650 headphones.

The spatialised sentences were presented simultaneously to the participants on each trial. The target sentences always used the word *Baron* as call sign, plus a random combination of colour and number. Maskers always used another call sign, colour and number, different to the ones used for the target. Participants were asked to listen to the target and select, among all the options, which colour-number combination they thought they heard in the target sentence.

Trials were grouped into blocks, and blocks into sessions. In each block, the target was always in the same position, which was displayed on the participant's interface. In this way, each block presents six different conditions defined by a combination of two HRTFs and three masker positions, one co-located and two separated. These conditions were repeated three times within each block, making a total of 18 trials per block, which were presented in a random order using Latin squares. The blocks were grouped in sets of three to form sessions, so that the three target positions were included in each session. There were a total of 6 sessions plus a training session at the beginning.

In the first block of the training session, only the target sentences were played to allow participants to familiarise themselves with the voices and sentence structure. The second and third blocks included target and masker sentences. All target positions were tested during training. Feedback was provided by highlighting the button displaying the correct answer. After the training session, no further feedback was available for the rest of the experiment.

2.3. HRTF-based objective metrics

Two objective metrics are computed in order to assess whether the differences in the behavioural assessment data between HRTF conditions can be explained by the differences in HRTF-specific numerical attributes. The first metric is based on spectral distortion (SD). This metric has been used multiple times to assess differences between HRTFs, as shown by Andreopoulou and Katz (2022). The computation starts by integrating the HRTF spectra at both ears into bands using ERB (equivalent rectangular bandwidth) bands to approximate human perception. The SD is defined as the root mean square of the difference between two HRTF spectra, thus quantifying how much one HRTF spectrum deviates from the other. The frequency range for the calculation is limited to the auditory frequency range (20 Hz – 20 kHz). The SD is computed at each ear for each target and masker HRTF spectra involved in the experiments. To obtain a single binaural value, the maximum (SD_{Max}) across ears is considered.

The second metric involved in the present study was the speech intelligibility model developed by Jelfs et al. (2011) to predict SRM from head-related impulse responses (HRIRs) that are used to spatialise the target and masker signals. This model was considered to explore its relevance predicting SRM differences between HRTF in the median plane. The model can account for binaural cues in the median plane due to HRTF asymmetry, as well as monaural spatial cues at each ear. The implementation available in the Auditory Modelling Toolbox was considered (Majdak et al., 2022; Lavandier et al., 2022), without any modifications. The HRIRs are passed through a gammatone filter bank to simulate the auditory frequency response. First, the better-ear SNR is estimated by comparing SNR across both ears. Second, the binaural unmasking advantage is computed following Culling et al. (2005). Finally, the values are combined and weighted across frequencies, giving more weight to the frequency bands relevant for speech. The output provides *effective* SNRs (Lavandier et al., 2022), which are relative values. Differences in this output directly predict differences in intelligibility scores (in dB). To model SRM, the model must be run twice: once for the co-located condition, using the same HRIR input for both the target and masker (this should result in a 0 dB *effective* SNR), and once for the separated condition, using distinct HRIRs for the target and masker. The output of the co-located condition is then subtracted from the output of the separated condition to yield a prediction of SRM.

The two metrics deviate from each other in major ways, making them both worth considering. First, the Jelfs model determines the better ear (the ear having the higher SNR) per frequency band, while the SD computation is based on the broadband differences across ears.

Second, due to the use of a root mean square, the SD computation leads to the same value when target and masker locations are swapped, which is not true for the Jelfs model. Finally, the Jelfs model also considers ITD in addition to the spectral aspects. Note that the two metrics are based only on the difference in HRTF spectra and thus do not consider any temporal variations of the speech signals. This is why another glimpsing model is proposed in Section 3.2.

2.4. Statistical analysis

SRM data were calculated by subtracting the intelligibility scores in the co-located condition from those in the separated conditions. The resulting values were then averaged across sessions and trials, to be further analysed using a linear mixed-effects model. Fixed effects in the model corresponded to the factors under investigation (refer to each experiment result section for further details, Sections 3.1 and 4.1). To account for individual variability, listener was included as a random intercept. The effect of individual HRTFs The normality of the model residuals was assessed using the Shapiro–Wilk test. The models were simplified using a stepwise backward elimination procedure. The analysis focused on the fixed-effect estimated coefficients, their confidence intervals, and associated t-tests, with a significance level set at 0.05.

The numerical metric values were used to assess whether the variance in SRM can be explained by the variance in HRTF-based numerical attributes by considering a correlation analysis. Individual SRM (SRM_{ind}) and differences in SRM between individual and KEMAR HRTF ($SRM_{Ind} - SRM_{Kem}$) were considered, as well as an average across spatial locations of these SRM, which were referred to as $AvSRM_{ind}$ and $AvSRM_{ind} - AvSRM_{Kem}$, respectively. The Native and Non-native groups were considered separately. The correlation analysis for each experiment computed 16 correlation coefficients (2 metrics × 2 native language × 4 SRM scores). A Bonferroni correction was used to ajust their $p - values$ for testing multiple (here, 16) hypotheses per experiment.

3. Experiment 1

Sixteen native English speakers (10 male, 6 female) with self-reported normal hearing participated in the test; all were staff or students at Imperial College London. The setup used to measure HRIRs was similar to that described by González-Toledo et al. (2024), which provided an exhaustive description; a brief summary is provided here. Participants were asked to place microphones in their ear canals (Knowles FG-23329-P07) and sit on a chair facing a pole with three loudspeakers (Genelec 8010A) at −50°, 0°, and 50°. The HRIRs were obtained by convolving the binaural recordings of a sine sweep with the sine sweep inverse signal. The measurement was performed in the room used to measure the SONICOM HRTF dataset as in Engel et al. (2023).

The signals were spatialised using either the individual HRTF or the KEMAR HRTF. For Experiment 1, the target was played simultaneously with one masker. The masker and target sentences were always uttered by different speakers of the same sex. They were simulated at three different elevations in the median plane (i.e., 0° azimuth): 50°, 0°, and −50°. This made a total of 18 conditions: 2 (HRTFs) × 3 (target positions) × 3 (masker positions).

3.1. Results

The intelligibility data collected in Experiment 1 are shown in Fig. 1 as blue symbols. The average percentages of correct words across participants were between 60% and 72% for the individual HRTF condition (left panel) and between 53% and 75% for the KEMAR HRTF condition (right panel). Increasing the target-masker spatial separation did not necessarily increase the percentage of correct answers. The Native group obtained higher intelligibility scores than the data collected with the Non-native group in González-Toledo et al. (2024); however, the benefit of increasing the spatial separation between target and masker was more pronounced for the Non-native group. The behaviour in the data collected by Martin et al. (2012) was quite different from the one obtained here, and this is further discussed in Section 5. The overall discrepancy with the existing data was the motivation for conducting Experiment 2.

The observations made above are further investigated with a regression model. The target location, masker location and HRTF were set as fixed effects. The null hypothesis of the Shapiro–Wilk normality tests was not rejected ($p = 0.27$); therefore linear mixed-effects models could indeed capture all the relevant effects. SRM was 10% higher when the masker was placed at 0° versus −50° ($\beta = 10.07$, CI = [3.86 16.28], $t = 3.38$, p = 0.002). The HRTF factor did not contribute significantly to the model fit, and was therefore excluded due to its negligible effect in this experimental setup. This was confirmed by the model parameters before stepwise regression, which indicated a decrease in mean SRM for the KEMAR HRTF, however, the effect size was low relative to the confidence intervals ($\beta = -1.38$, CI = [−22.14 19.36], $t = -0.12$, p = 0.89).

The results of the correlation analysis between numerical metrics and SRM (see Table 1, top table) differ between the two groups, probably because the effect of spatial separation between target and masker was not significant for the Native group. The only significant correlation was between Individual SRM and the Jelfs SRM for the Non-native group, when considering the scores averaged across spatial locations. A negative coefficient was reported, which means that lower predicted SRM (Jelfs model) have higher measured SRM—an unexpected outcome regarding EM, assuming that SRM would increase if HRTF-based spectral cues were more important. None of the correlations involving differences between HRTF conditions were significant. The associated linear regressions are plotted in the four left panels of Fig. 2 as solid lines in blue or orange for the Native or Non-native groups, respectively. The top panels depict the data obtained with SRM scores concatenated across spatial positions, while the bottom panels show the scores averaged across spatial conditions. In the second top panel (individual SRM vs. Jelfs model), the slopes of the linear regressions are negative, consistent with the negative correlation coefficients reported in Table 1. The data points show considerable dispersion, limiting the strength of any conclusions regarding the relationship between the measured and predicted SRM values.

3.2. Glimpsing analysis

The difference in intelligibility data observed between the Native and Non-native groups suggests that the two groups used different strategies to complete the task. While the Non-native group benefited from the spatial separations between target and masker, the Native group performed similarly regardless of the masker and target placement. This may indicate that the Native group relied on spectral glimpsing, integrating speech features across time and frequency. Such a behavioural difference may be due to the increased task difficulty caused by the use of a foreign speech corpus. To investigate this further, a glimpsing analysis was carried out to assess whether the Native group indeed relied more on this strategy than the Non-native group. Such models provide accurate intelligibility predictions (e.g., Wasiuk et al., 2023; Edraki et al., 2022; Best et al., 2017). The proposed implementation was inspired by Wasiuk et al. (2023). The glimpsing analysis conducted here was designed to investigate whether intelligibility scores increased with target glimpsing, and how this interacts with spatial separation. In this way, it serves more as a descriptive tool rather than a predictive model, meaning that we are not employing target glimpsing to be directlyconverted into percentages of correct answers. Instead, it was used as a means to further analyse the data.

Only the monaural versions of the sentences were used for the analysis, in order to isolate the effect of spectral differences between different talkers. The analysis consisted of four main stages:

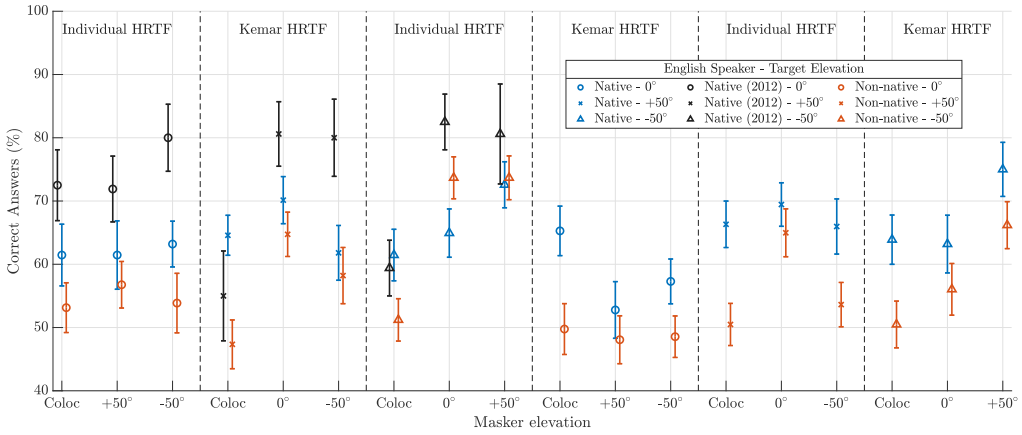

Fig. 1. Average percentage of correct answers measured in Experiment 1. The figure is divided into six panels, representing the combinations of HRTF and target elevation conditions. The masker elevations vary along the abscissa. The blue symbols depict the data collected for the current study while the orange ones are the data collected by the University of Malaga with the Non-native group in the previous study (González-Toledo et al., 2024). For sake of comparison, the data of Martin et al. (2012) are also displayed, as black symbols. The different marker symbols are showing different target elevations (circles for 0°, crosses for +50°, and upward triangles for −50°).

Table 1
Results of the correlation analysis between numerical metrics and behavioural data of Experiment 1 (top table) and Experiment 2 (bottom table). The correlation analysis was conducted separately for the Native and Non-native groups. Each correlation was computed between a numerical metric and SRM data. The corresponding p-value of each coefficients are represented as * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$) or **** ($p \leq 0.0001$). A Bonferroni correction adjusted the p-value to the 16 tested hypotheses per experiment.

	SRM_{Ind}	$SRM_{Ind} - SRM_{Kem}$	$AvSRM_{Ind}$	$AvSRM_{Ind} - AvSRM_{Kem}$
Exp 1				
Natives				
MaxSD	−0.02	0.04	−0.06	0.04
Jelfs	−0.21	0.03	−0.24	−0.11
Non-natives				
MaxSD	0.18	−0.05	0.14	0.25
Jelfs	−0.32**	−0.06	0.43	−0.03
Exp 2				
Natives				
MaxSD	0.61	0.00	0.35	0.19
Jelfs	0.83****	0.24	0.95***	0.45
Non-natives				
MaxSD	0.28	−0.16	−0.09	−0.16
Jelfs	0.46	0.02	0.15	−0.06

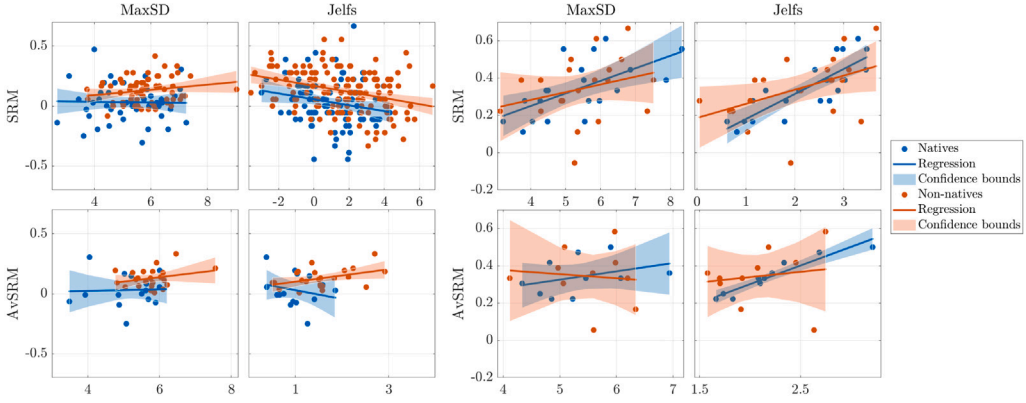

Fig. 2. SRM scores as a function of the objective metrics for Experiment 1 (four left panels) and Experiment 2 (four right panels). Only the SRM obtained with individual HRTF are depicted here. The top panels show the SRM scores concatenated across spatial conditions, and the bottom panels the SRM average across spatial conditions. The solid lines represent the linear regressions and the shaded area the 95% confidence bounds. The same colour pattern as in Fig. 1 is used to differentiate the different groups.

- pre-selection of the target-masker sentence pairs included in the study;
- short-term spectral analysis of each pair to derive a spectral glimpsing value;
- classification of the sentences into 5 groups based on the quintiles of their spectral glimpsing values;
- classification of the participants' answers according to the sentence pairs used in the trials.

The pre-selection of sentences consisted in reducing the number of sentence combinations to only the ones involved in the experiment. Target sentences were those using "baron" as the call sign. Hence, the total number of target sentences was 256 (8 target digits × 4 target colours × 8 speakers). Masking sentences included all those using any of the remaining call signs. Additionally, each masker sentence had to be uttered by a same-sex speaker, and both the colour and digit had to differ from those of the target. For each target sentence, 441 masking sentences were possible (3 other same-sex speakers × 7 call signs × 7 digits × 3 colours). This resulted in 112,896 target-masker sentence combinations to analyse.

The short-term spectral analysis started with framing the target and masking sentences using a 12 ms square window without overlap. Next, we derived the short-term spectra using a gammatone filterbank. We finally computed and integrated the absolute SNRs for each band, applying a Speech Intelligibility Index (SII) weighting. This process allowed us to obtain a broadband value for absolute spectral glimpsing. Absolute SNRs were considered because, in scenarios with two competing talkers, absolute intensity differences matter more than relative ones as argued by Brungart (2001). The resulting spectral glimpsing values were collected across all target-masker sentence pairs and grouped into five quintiles. Thus, target-masker sentences were categorised into 5 groups.

Finally, participants' answers were grouped according to the spectral glimpsing classification of the sentences and averaged to obtain a percentage of correct answers for each quintile. The method was applied to all participants' answers (Coloc + Separ), as well as separately for those obtained in the co-located condition (Coloc-only), and the separated condition (Separ-only).

Ideally, the percentage of correct answers should increase with higher spectral glimpsing quintiles, as more cues become available. Moreover, if participants relied on spectral glimpsing rather than spatially induced spectral cues, the Coloc+Separ, Coloc-only and Separ-only curves would overlap. Conversely, if spatial separation served as the main cue for participants, the Coloc-only and Separ-only curves would fall below and above the Coloc+Separ curve, respectively.

The results of the analysis are shown in Fig. 3. The percentages of correct answers increase with spectral glimpsing (left panel), indicating that both Native (blue) and Non-native (orange) groups used spectral differences between voices to unmask the target. The effects of spectral glimpsing on behavioural data are approximately 40% for both groups.

The dashed lines in the left panel represent the average answers obtained in the co-located condition (pentagrams) and the separated spatial condition (downward triangles). The differences between these lines reflect the SRM, displayed in the right panel. The benefits of spectral glimpsing and spatial cues seem to depend on the listeners' native language.

To confirm those observations, a linear mixed-effects model was designed to analyse the percentage of correct answers. The factor listener was set as random effect, while source spatial separation (Co-located or Separated), HRTF (Individual or KEMAR), native language (Natives or Non-natives), and spectral glimpsing (five spectral glimpsing quintiles) were set as fixed effects. Since the residuals of this model were not normally distributed (Shapiro–Wilk test, $p = 0.004$), a Box–Cox transformation was applied to identify the optimal λ that maximises the normality of model residuals (here, $\lambda = 1.71$). A stepwise regression similar to the method described in Section 2.4 was performed to simplify the model.

A summary of the model parameters is presented in Table 2. The variation in the percentage of correct answers is primarily driven by spectral glimpsing, which shows an increase of approximately 52% between the first and last quintiles ($\beta = 51.80$, CI = [47.72, 55.89], $t = 24.77$, $p < 0.0001$). The second most notable effect is native language, as the Non-native group exhibited lower intelligibility scores ($\beta = -13.58$, CI = [−25.35, −1.82], $t = -2.26$, $p = 0.03$). Intelligibility scores were higher in the separated conditions ($\beta = 5.67$, CI = [0.97, 10.37], $t = 2.36$, $p = 0.019$). However, the interaction between the separated condition and the KEMAR HRTF degraded intelligibility ($\beta = -7.46$, CI = [−12.62, −2.29], $t = -2.82$, $p = 0.005$). This effect can be explained by the integration of the dataset from (González-Toledo et al., 2024) into the present analysis, as their findings indicated an effect of HRTF on SRM.

4. Experiment 2

Based on the findings from Experiment 1, a second experiment was conducted. The rationale was to increase task difficulty by adding a second masking sentence, thus decreasing spectral glimpsing and allowing for a deeper investigation of behavioural strategies in a speech-on-speech paradigm. In particular, this aimed to provide insight into the relationship between task difficulty (as varied by language proficiency and number of maskers) and SRM. This experiment reproduces some of the conditions reported by Best (2004).

The process for measuring HRTFs in Experiment 2 was similar to that of Experiment 1, except for changes in the measurement setup. The SONICOM HRTF setup (Engel et al., 2023) was used instead, as participants were recruited only at Imperial College London and could attend the measurement session in the lab. The procedure was shortened, since only positions in the median plane were of interest for this experiment.

Ten natives English speakers[1] (7 males, 2 females, 1 prefer not to say) and ten non-native English speakers (7 males, 3 females) subjects, all with self-reported normal hearing, took part in Experiment 2. They were all students or staff members from Imperial College London. The non-native English speakers had spent on average 7 years in an English-speaking country, making the results broadly comparable to those of González-Toledo et al. (2024), where the average certified English level was B2 ("[...] a person can understand the main ideas of complex texts and can interact with native speakers with fluency and spontaneity".).

In order to increase the difficulty of Experiment 2, the simplest and most effective adjustment was considered to be the increase of the number of maskers to two, reducing their individual level by −3 dB. This allowed for the SNR to remain unchanged, but for the task difficulty to increase due to informational masking (Freyman et al., 2004). The two maskers were always co-located to each other, resulting in a two-talker masker. The masking sentences were always uttered by a different speaker than the target; and the speakers' sex could differ. The target was always simulated in front of the listener (0° azimuth, 0° elevation) while the two maskers were simulated also frontally, at three elevations (45°, 0°, and −45°). This generated a total of 6 conditions (2 HRTFs × 1 target position × 3 masker positions). It is important to underline again that the sole difference between experimental procedures in Experiment 2 and 1 was due to the addition of the second masker and reduction of the individual level of −3 dB.

[1] One also took part in Experiment 1.

Table 2
Parameter estimates from the linear mixed-effects model assessing the fixed effects of tarket-masker location (Co-located or Separated), native language (Natives or Non-natives), HRTF (Individual or KEMAR), spectral glimpsing quintiles (SGq1, SGq2, SGq3, SGq4, SGq5), and interactions on the percentage of correct answers for Experiment 1. The listener factor is set as random effect. SE stands for standard error; df for degrees of freedom; CI for confidence interval at 95%.

Effect	Estimate (β)	SE	df	t-value	p-value	CI (95%)
(Intercept)	50.39	4.82	49	10.46	<0.0001	[41.00, 59.78]
Separated	5.67	2.41	695	2.36	0.019	[0.97, 10.37]
Non-natives	−13.58	6.02	39	−2.26	0.030	[−25.35, −1.82]
SGq2	10.41	2.09	695	4.98	<0.0001	[6.33, 14.49]
SGq3	31.47	2.09	695	14.57	<0.0001	[26.39, 34.55]
SGq4	43.89	2.09	695	20.98	<0.0001	[39.80, 47.97]
SGq5	51.80	2.09	695	24.77	<0.0001	[47.72, 55.89]
KEMAR	2.18	1.87	695	1.16	0.25	[−1.47,5.83]
Separated x Non-natives	7.71	2.67	695	2.89	0.004	[2.50, 12.92]
Separated x KEMAR	−7.46	2.65	695	−2.82	0.005	[−12.62, −2.29]

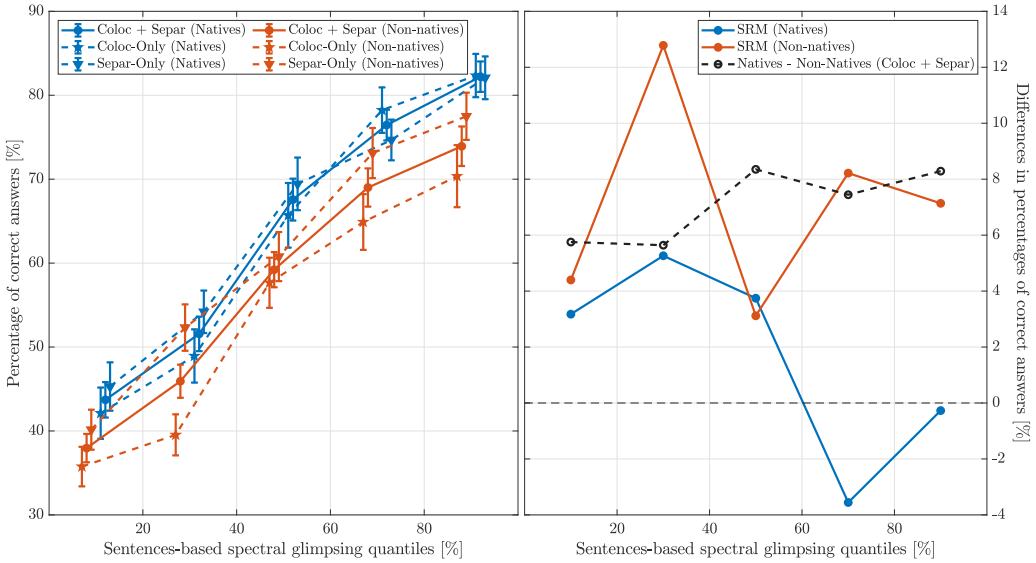

Fig. 3. Left panel: Percentage of correct answers (with standard errors) plotted as a function of spectral glimpsing quintiles. Solid lines represent the percentages obtained considering both co-located and separated conditions. Dotted lines indicate the percentages for the co-located setup (pentagram symbols) and the separated setup (downward triangles). Blue corresponds to Native results and Orange to Non-native results. Right panel: Differences in percentages of correct answers as a function of spectral glimpsing quintiles. The black dotted line represents the difference between the Native and Non-native groups across spatial conditions. Differences in intelligibility between separated and co-located setups — i.e. SRM scores — are shown in solid lines using the same colour scheme as in the left panel.

4.1. Results

The results of Experiment 2 are displayed in Fig. 4. The effect of the spatial separation between target and maskers was larger than in Experiment 1, which can be attributed to the type of masking induced by the two-talker masker involved in Experiment 2. This addition substantially decreased the percentage of correct answers for the co-located conditions (from about 60% in Experiment 1, to 35% in Experiment 2), while percentages for the separated conditions remained roughly unchanged. The SRM, expressed as the difference in percentage of correct answers between the co-located and separated conditions, reached 40% in some cases. Intelligibility scores were higher when sources were simulated with individual HRTFs.

On average, across listening conditions, the Non-native group performed similarly to the Native group, unlike in Experiment 1. This convergence likely reflects that, the Non-native group showed better English proficiency than those in González-Toledo et al. (2024), or the two-talker masker mitigated group-related performance differences by promoting a stronger reliance on available spatial cues and spectral glimpsing. These experimental conditions replicate those in Best (2004), and the current perceptual data led to intelligibility scores similar to those reported in the original study.

The qualitative observations were further examined using a linear mixed-effects model to analyse SRM data. The fixed effects considered were HRTF, native language, and masker location. The results of the Shapiro–Wilk test ($p = 0.26$) suggest that all the effects can be captured by linear mixed-effects models. The mean SRM was 6.9% lower for the KEMAR HRTFs than for the individual HRTFs ($\beta = -6.94$, CI = [−12.13–1.76], $t = -2.62$, p = 0.01); and 15.6% lower when the masker was placed at +45° rather than −45° ($\beta = -15.56$, CI = [−20.74–10.37], $t = -5.87$, $p < 0.0001$). The native language factor was excluded from the final model during the stepwise backward elimination procedure. This suggests that, under Experiment 2 conditions, the effect of native language was not statistically meaningful. This model before simplification indicated that the Non-native group showed a mean SRM slightly lower than the Native group but this was marginal compared with the confidence interval ($\beta = -0.55$, CI = [−12.37 11.26], $t = -0.09$, p = 0.93).

The correlation analysis (see Table 1, bottom table) highlights a few relevant differences between the Native and Non-native groups for the Individual HRTF condition. The two HRTF-based metrics considered in the analysis were significantly correlated with SRM for the Native group, whereas for the Non-native group no correlation coefficients were significant after Bonferroni correction. None of the correlations involving differences between HRTF conditions reached significance.

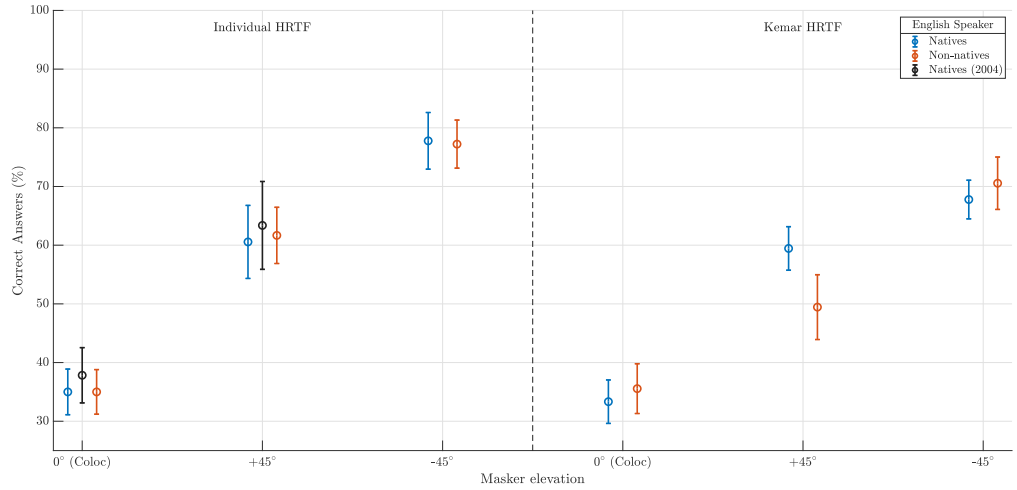

Fig. 4. Intelligibility scores collected in Experiment 2. The figure is split into 2 panels for individual (left-hand side) and KEMAR (right-hand side) HRTF conditions. The masker elevations changed along the abscissa axis. The blue and orange symbols depict the data collected with the Native and Non-native groups, respectively. For sake of comparison, the data of Best (2004) are also displayed, as black symbols.

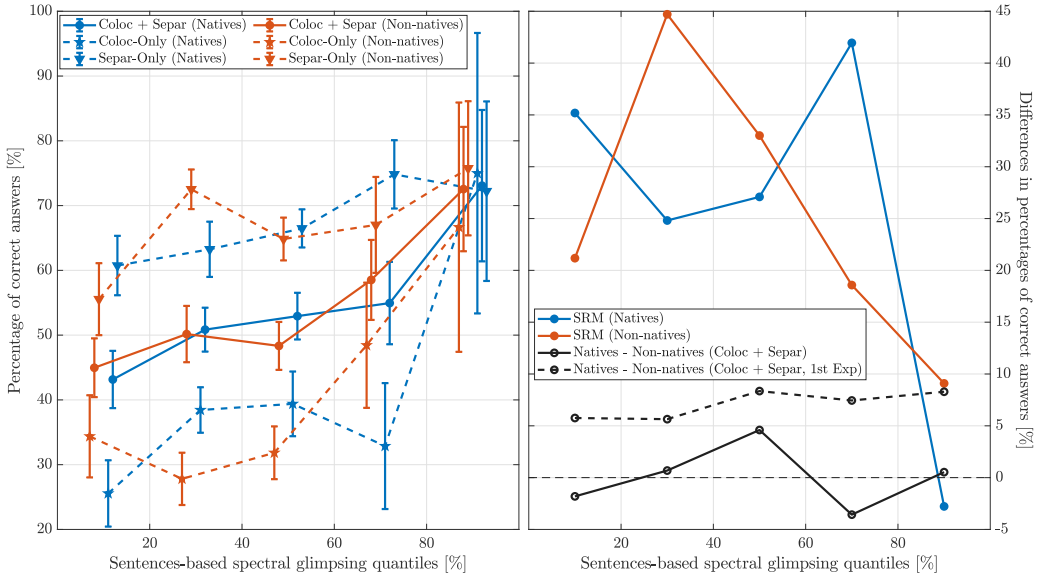

Fig. 5. Left panel: Percentage of correct answers (with standard errors) plotted as a function of spectral glimpsing quintiles. Solid lines represent the percentages obtained considering both co-located and separated conditions. Dotted lines indicate results from the co-located setup (pentagram symbols) and the separated setup (downward triangles). Blue corresponds to the Native group, and orange to the Non-native group. Right panel: Differences in percentage of correct answers are plotted as a function of spectral glimpsing quintiles. The differences in intelligibility between the separated and co-located setups (i.e., SRM scores) are shown in solid lines following the same colour scheme as in the left panel. The solid black line shows the difference between the Native and Non-native groups across spatial conditions. For comparison, the difference between the Native scores in Experiment 1 and the Non-native scores reported in González-Toledo et al. (2024) is also shown (dotted black line).

In Fig. 2, the associated linear regressions are plotted in the four right panels. The confidence bounds for the Native group are narrower than for the Non-native group, which indicates more uncertainty in the regression for the latter.

4.2. Glimpsing analysis

The sentence-based spectral glimpsing quintiles were computed following the same steps as in Experiment 1 (see Section 3.2), using the spectra of target and masker sentences defined according to the criteria of Experiment 2 (see Sections 2 and 4). Spectral glimpsing was computed by integrating the absolute difference between the target spectrum and the combined masker spectrum, using a SII weighting. The resulting values were then grouped into quintiles.

The answers of each trial of Experiment 2 (excluding the training trials) were subsequently categorised using the combinations of target and masker sentences, along with the spatial conditions and native languages. If a given combination corresponded to the nth spectral glimpsing quintile, the answer was assigned to the nth category. The mean percentages for these categories are displayed in Fig. 5.

In the left panel, data corresponding to the Native and Non-native groups are shown in blue and orange, respectively. Comparing the dashed lines with pentagrams (co-located condition) to those with triangles (separated condition) suggests that spatial cues were irrelevant when the target glimpsing was high. This is further confirmed by the coloured lines in the right panel, which depict SRM. These results suggest that both groups relied predominantly on spatially induced spectral cues, while spectral glimpsing contributed little, except glimpsing was sufficiently large—an effect not observed in Experiment 1.

Table 3
Parameter estimates from the linear mixed-effects model assessing the fixed effects of tarket-masker location (Co-located or Separated), native language (Natives or Non-natives), HRTF (Individual or KEMAR), spectral glimpsing quintiles (SGq1, SGq2, SGq3, SGq4 or SGq5), and interactions on the percentage of correct answers for Experiment 2. The listener factor is set as random effect. SE stands for standard error; df for degrees of freedom; CI for confidence interval at 95%.

Effect	Estimate	SE	df	*t*-value	*p*-value	CI (95%)
(Intercept)	25.70	7.37	224	3.49	0.0006	[11.86, 39.54]
Separated	39.50	9.70	295	4.07	0.0001	[21.28, 57.72]
SGq5	35.10	15.06	300	2.33	0.02	[6.84, 63.46]
SGq5:KEMAR	50.76	19.31	302	2.63	0.009	[14.31, 87.00]
Separated x Non-natives x SGq2	33.90	15.83	295	2.14	0.03	[4.16, 63.65]
Separated x SGq5 x KEMAR	−62.50	23.75	301	−2.63	0.009	[−107.07, −17.74]

Similarly to Experiment 1, a linear mixed-effects model was employed. The factor listener was set as a random effect, while source spatial separation (Co-located or Ceparated), HRTF (Individual vs. KEMAR), native language (Natives or Non-natives), and spectral glimpsing (5 spectral glimpsing quintiles) were set as fixed effects. The normality of the residuals was verified.

A summary of the model parameters with significant effect is presented in Table 3. The variations in the percentage of correct answers were about the same for the separated condition (β = 39.50, CI = [21.28, 57.72], t = 4.07, p = 0.0001) and SGq5 (β = 35.10, CI = [6.84, 63.46], t = 2.33, p = 0.02). On one hand, this supports measurement of SRM and suggests that only high spectral glimpsing improved prediction scores. There was a significant interaction between SGq5 and KEMAR HRTF (β = 50.76, CI = [14.31, 87.00], t = 2.629, p = 0.009). Moreover, a significant three-way interaction was reported between the separated condition, the fifth quintile, and KEMAR HRTF (β = −62.50, CI = [−107.07,-17.], t = −2.63, p = 0.009). This indicates that the effect of SGq5 on intelligibility was significantly stronger when combined with KEMAR HRTF compared to individual HRTF, but the effect of spatial separation was significantly reduced when combined with SGq5 and KEMAR HRTF. Both observations together suggest that individual HRTFs provided further improvement for the separated condition when the trial presented high spectral glimpsing (SGq5). There was a significant three-way interaction between the Separated condition, the Non-native group, and SGq2 (β = 33.90, CI = [4.16, 63.65], t = 2.14, p = 0.033), i.e., the effect of spatial separation was not consistent across all conditions. It was specifically enhanced for the Non-native group in the second glimpsing quintile condition.

5. General discussions

5.1. Comparison with the replicated experiments

Experiment 1 was conducted as a replication of the study by González-Toledo et al. (2024) but recruiting a Native group. The study by González-Toledo et al. (2024) was itself conceptually based on the framework proposed by Martin et al. (2012). Thus, Experiment 1 also builds on that conceptual basis while replicating the procedures described in González-Toledo et al. (2024). More specifically, in Martin et al. (2012) the SRM was measured using the same source positions as the ones employed in Experiments 1, as well as different HRTFs conditions, including the individual participants' HRTFs. The main difference between Martin et al. (2012) and González-Toledo et al. (2024) was the participants, in the first case the Native group, in the second case not. This is possibly the main reason why the observed SRM was 17% in the first and 12.5% in the second, indicating that higher language proficiency resulted in a smaller SRM advantage. Furthermore, Martin et al. (2012) normalised the SNR at 0 dB at each ear, while in González-Toledo et al. (2024) the SNR was normalised at 0 dB across ears. This proposed normalisation could have been beneficial, as it provides an ear with a better SNR. In doing so, the degradation due to the native language condition when comparing the data from the two experiments could have been underestimated.

Regarding the Native group in Experiment 1, HRTF individualisation was not reported as a significant factor for SRM, which did not correspond to what was found in González-Toledo et al. (2024) with the Non-native group. Moreover, the percentages of correct words for any conditions were about the same as for the co-located condition in Martin et al. (2012). This was a surprising result that led to further investigations. Experimental differences (including software and hardware) were generally discarded as an explanation for the discrepancies, because the exact same application, HRTF measurement setup and signal processing chain were used in both experiments and locations. This absence of logical explanations was the main motivation to design and run Experiment 2, which was based on Best (2004), involving conditions similar to Experiment 1 but with a two-talker masker instead of one. There was also the opportunity to maximise the chance of measuring SRM by increasing the number of spatially co-located speech maskers (Hawley et al., 2004; Freyman et al., 2004). The percentages of correct words and SRMs measured in Experiment 2 were similar to the data from the replicated study by González-Toledo et al. (2024). This suggested again the absence of hardware or software problems. As clear from the results presented above and discussed later on, the addition of a masker seems to provide an explanation for the discrepancy between the results of González-Toledo et al. (2024) and Experiment 1, but not yet for the differences between Experiment 1 and Martin et al. (2012), where an SRM advantage was indeed found for the Native group and employing one single masker. Although beyond the scope of this work, further explorations are needed to explain these discrepancies.

5.2. On the interaction between individual HRTF and task difficulty

The present study manipulated task difficulty by introducing a second speech masker in Experiment 2 and by recruiting participants with varying degrees of English proficiency (Natives and Non-natives). These two factors introduce variations in EM and IM that may interact between each other as well as with HRTF individualisation. The current section aims to review these effects and compare them to the present dataset and the dataset from González-Toledo et al. (2024).

The amounts of EM and IM increased with the introduction of a second masker in Experiment 2. In both experiments, the long-term broadband SNR was maintained at 0 dB; however, the spectro-temporal overlaps between the target and the two-talker maskers also increased, likely resulting in higher short-term, band-specific EM. Additionally, when the masker comprises multiple speech signals, IM is maximised when considering a two-talker masker (Freyman et al., 2004). SRMs scores in González-Toledo et al. (2024) and Experiment 1 were much lower compared to those measured in Experiment 2. This proves that the amount of masking (IM and EM) increased due to the two-talker masker in Experiment 2. Moreover, Non-natives are more affected by IM (Garcia Lecumberri et al., 2010) than Natives; however, this difference is likely reduced as language proficiency increases. In Experiment 2, the Non-native group was recruited in an English-speaking country, and was likely more proficient in English than the group in González-Toledo et al. (2024), who lived in a Spanish-speaking country. Therefore, the task may have been easier for the Non-native group in Experiment 2. Altogether, this suggests that IM and EM increased

in Experiment 2 due to the two-talker masker, while IM decreased for the Non-native group due to their improved language proficiency. This was confirmed by the model designed for the glimpsing analysis. In Experiment 1, the Non-native group obtained lower intelligibility scores (see Table 2), but this trend was not observed in Experiment 2 (Non-native effect: $\beta = 8.81$, CI = [−7.71; 25.3], $t = 1.00$, $p = 0.32$). In conclusion, the improved English proficiency of the Non-native group in Experiment 2 compensated for the increase in masking, achieving intelligibility scores similar to the Native group. Of course our binary categorisation of language proficiency does not allow to properly characterise the relationship between these factors, but in this context it would have been complex to use a more granular definition, and would not have allowed us to reach sufficient statistical power. Further studies are therefore needed to better clarify the matter.

These variations in EM and IM may interact with the effects of HRTFs. Individualising HRTFs potentially reduces IM by enhancing the spatial rendering of the sound sources (Freyman et al., 1999, 2004), thereby improving stream segregation. Furthermore, each HRTF provides an intrinsic amount of spatially induced spectral cues that facilitate release from EM. The interaction between the effect of HRTF individualisation and the one of language proficiency on IM can explain the discrepancy between Experiment 1 and the data from González-Toledo et al. (2024). The Non-native group recruited by González-Toledo et al. (2024) faced more IM due to language proficiency, which increased IM, thus decreasing their percentage of correct words. However, the integration of individual HRTFs allowed to provide cues that compensate for this increased IM. This is plausible explanation of the significant effect of HRTF individualisation on SRM reported by González-Toledo et al. (2024) but not in Experiment 1. In Experiment 2, linear mixed-effects model developed to analyse SRM data reported an effect size of the interaction between HRTF and native language that was rather low as opposed to the confidence interval ($\beta = 1.11$, CI = [−12.99, 15.21], $t = 0.149$, $p = 0.88$), which means that both groups were similarly affected by HRTF individualisation, and then the Non-native group did not experience higher IM due to language proficiency.

Regarding the interaction between the effect of HRTF individualisation and type of maskers, in Experiment 2 an approximate increase in SRM of 6.9% (on average across groups) was measured with individual HRTFs as opposed to the SRM collected with the KEMAR HRTF. A similar trend was observed in González-Toledo et al. (2024). These improvements could not be explained by differences between objective HRTF-based metrics for KEMAR and individual HRTF, suggesting that the improvement can be due to the fact that employing individual HRTF provides a more consistent perceived source position, which might improve SRM by increasing the release from IM. Another explanation could that participants were unable to efficiently employ the unfamiliar spatial cues in the KEMAR HRTF.[2] This effect was not observed for the Native group in Experiment 1, and this can be explained by the fact they relied on spectral glimpsing instead of spatially induced spectral cues. This supports the idea that the benefit of familiar spatial cues cannot be assessed by a simple HRTF-based energetic model in this setup, likely because IM is involved.

The correlation coefficients in Experiment 2 were notably high for the Native group when comparing individual SRM scores with the Jelfs model output, whose predictions explained at least 69% ($r^2 \geq 0.69$) of the SRM variance. This suggests that individuals with HRTFs providing more cues (due to their morphology) benefit from improved SRM. In other words, HRTFs equally well-suited to each listener (e.g., individualised HRTFs) provide intrinsic release from EM, while the release from IM remains similar (and is likely maximised). The variation in SRM observed when using non-individual HRTFs (in this case, KEMAR) is therefore related to each participant's ability to utilise non-individual cues, rather than the actual differences in objective HRTF cues.

5.3. Objective metrics

Two kinds of objective metric were used to analyse the data: (1) two HRTF-based metrics to assess the differences between HRTFs or source locations (see Section 2.3); and (2) a sentence-based metric evaluating glimpsing (see Sections 3.2 and 4.2). The HRTF-based metrics were correlated with the SRM data to evaluate the participants' ability to use both spatially induced and intrinsic spectral cues to improve intelligibility. This is a way to evaluate the spatial release from EM that is inherent to the HRTFs, and to advance on the development of a model/method to numerically quantify perceptually-relevant differences between HRTFs. On the other hand, the glimpsing model is more adapted to the conditions involving speech maskers, because it considers target audibility and spectro-temporal signal modulations (e.g., Wasiuk et al., 2022, 2023; Best et al., 2017).

In Experiment 1, MaxSD was not significantly correlated with SRM. The Jelfs model outputs, however, showed a negative correlation with SRM when data were concatenated across spatial conditions for the Non-native group (González-Toledo et al., 2024). This result was surprising, as it contradicted the expected effect of release from EM. Nevertheless, Brungart (2001) demonstrated that, in a two-competitive-talker paradigm, decreasing the target level below 0 dB SNR does not necessarily impair intelligibility. The most likely explanation for this effect is that listeners can segregate two concurrent speech sources based on level differences. Thus, the *a priori* negative effect of decreasing the SNR on intelligibility (i.e, increased EM) may be offset by improved source segregation (i.e., reduced IM). It is noteworthy, however, that this conclusion should be interpreted with caution, as only one correlation coefficient was significant, and its squared value was low ($r^2 \leq 0.10$), meaning that the effect of HRTF-based cue differences between source locations was likely secondary for this task.

In Experiment 2, MaxSD was not significantly correlated with SRM. The Jelfs model predictions were better correlated with the individual SRM for the Native group. The computation of the Jelfs model chooses the better-ear per frequency bands and considers relative values in SNR, while the MaxSD metric relies on the absolute difference in SNR and selects the better-ear based on broadband values (see Section 2.3). In other words, this exhibits the importance of relative SNRs and/or frequency-dependent better ear in the individual HRTF condition in Experiment 2.

The glimpsing model was computed to investigate the relationship between glimpsing, spatial cues, IM, and native language. In the analysis of the data from Experiment 1 and in González-Toledo et al. (2024), the variation in intelligibility due to spectral glimpsing was estimated to be about 52% (increase from the lowest to the highest quintiles). Meanwhile, the effect of source location was approximately 6% for the Native and Non-native groups, with an additional improvement of 8% for the Non-native group in González-Toledo et al. (2024), as shown by Table 2. Thus, speech intelligibility was primarily driven by the intrinsic spectral differences between the target and masker voices (i.e., spectral glimpsing), while the additional spectral differences introduced by the HRTF with changes in elevation (i.e., spatially induced cues) had only a minor effect. In other words, the spectral distinctiveness of the original speech recordings dominated over the spectral modifications produced by the source location in this experimental setup.

In Experiment 2, which involved a two-talker masker, increasing spectral glimpsing improved speech intelligibility, and this improvement appeared to benefit both Native and Non-native groups similarly (black solid line, right panel, Fig. 5). Furthermore, the dashed lines in the left panel may reflect an interaction between the influence of spectral glimpsing and spatial conditions. This is confirmed by the

[2] It might well be that the KEMAR HRTF is in general of lower quality compared to the individual ones. However, the KEMAR HRTF was measured in the same lab and using the same setup used as the individual HRTFs, and it has been extensively employed in prior spatial hearing research as a *generic* filter, making it a suitable candidate for comparison in this study.

linear mixed-effects model presented in Table 3 only for the Non-native group, which means it successfully combined spatial cues and spectral glimpsing to solve the task.

Furthermore, The co-located curves (dashed lines with pentagrams) rise substantially at the highest level of spectral glimpsing, indicating that participants better understood the target when there was a relatively high spectral difference between the target and maskers. Conversely, this increase was not observed for the separated conditions (dashed lines with downward triangles), or was observed to a lesser extent. This suggests that the primary cue in this experiment was the spatial condition. This is partially confirmed by Table 3, which indicates a slightly larger effect size for the separated condition compared to the fifth spectral glimpsing quintile. However, this conclusion can be only partially validated as their confidence intervals substantially overlap. These observations align with the conclusions of Wasiuk et al. (2023), whose experimental design is comparable to that of Experiment 2. Wasiuk et al. measured speech recognition involving a two-talker masker, which was either co-located with the target or spatially separated along the horizontal plane. They concluded that the amount of glimpsing required to understand the target is larger in the co-located spatial condition. This is likely because IM was maximised in the co-located condition and substantially reduced in the separated condition, as spatial cues improved target segregation. The present study extends this conclusion to the median plane.

An opposite pattern can be observed in the use of spectral glimpsing and spatially induced spectral cues when comparing the two experiments. In Experiment 1, participants relied more on spectral glimpsing, as the primary increase in the percentage of correct answers was driven by an increase in spectral glimpsing (in Fig. 3, the increases in the solid lines are larger than the differences between the dashed lines). In contrast, in Experiment 2, the benefit of spatially separating the target source from the masking sources was slightly larger than the benefit of increasing spectral differences between the target and the two-talker maskers (in Fig. 5, the increases in the solid lines are shallower compared to Experiment 1, while the differences between the dashed lines are larger). This is likely due to the effects of the two-talker maskers in Experiment 2, which increased both IM and EM. Furthermore, the fact that the masker contained two talkers may have enhanced intelligibility in the separated condition, as the masking voices might have partially masked each other.

6. Conclusion

The present study investigated the strategies employed by participants to solve a speech-on-speech task involving sources spatialised in the median plane. Task difficulty was manipulated by varying the number of speech maskers and the native language of the participants. Additionally, sources were spatialised using either individual HRTFs or a generic KEMAR HRTF.

To analyse participants' strategies, two HRTF-based metrics and a glimpsing model were used. The HRTF-based metrics assessed spatial and spectral cues, while the glimpsing model evaluated target glimpses independently of spatial location. Results indicate that participants adopted different strategies to solve the task, suggesting a potential trade-off between relying on spectro-temporal glimpsing and HRTF-related spatial spectral cues. In Experiment 1 (one competing speech masker), the Native group primarily relied on target glimpsing, whereas the Non-native group relied more on spatial cues and showed a lower percentage of correct responses. This suggests that the Non-native group was less able to benefit from target glimpsing and compensated by using spatial information. In Experiment 2 (two-talker maskers), both groups relied on spatial cues and spectral glimpsing. The Non-native successfully combined both cues to further improve intelligibility—a trend not observed for the Native group. Overall, the results suggest that spatial cues become increasingly important as task difficulty increases (due to non-native language status or a higher number of maskers), whereas the benefit of target glimpsing decreases under more challenging conditions. Moreover, the comparison between individual and generic HRTFs showed that individualised spectral cues further enhanced performance in difficult listening situations.

In order to better explain and understand the behavioural outcomes, experiments measuring higher-level perceptual mechanisms (e.g. listening effort and cognitive load) could be designed and carried out in the future. At the same time, objective metrics such as the ones presented here have been used in the past in related research (e.g. Cuevas-Rodriguez et al., 2021), and allow not only to partly explain some of the observed mechanisms, but also to directly compare our results with the ones of previous studies.

CRediT authorship contribution statement

Thibault Vicente: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Daniel González-Toledo:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Conceptualization. **María Cuevas-Rodríguez:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Luis Molina-Tanco:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Arcadio Reyes-Lecuona:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Conceptualization. **Lorenzo Picinali:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Conceptualization.

Acknowledgements

We would like to thank Dr. Virgina Ann Best for providing a copy of the CRM corpus, which has been used in this experiment. This study has been supported by SONICOM (www.sonicom.eu), a project funded by the European Union's Horizon 2020 research and innovation program under grant agreement No. 101017743, and the Spanish National Projects SAVLab and SONIX, under grants No. PID2019-107854GB-I00 and PID2023-152547NB-I00 respectively.

References

Andreopoulou, A., Katz, B.F., 2022. Perceptual impact on localization quality evaluations of common pre-processing for non-individual head-related transfer functions. AES: J. Audio Eng. Soc. 70 (5), 340–354. http://dx.doi.org/10.17743/jaes.2022.0008.

Berwick, N., Lee, H., 2020. Spatial unmasking effect on speech reception threshold in the median plane. Appl. Sci. (Switzerland) 10 (15), http://dx.doi.org/10.3390/APP10155257.

Best, V., 2004. Spatial Hearing with Simultaneous Sound Sources : A Psychophysical Investigation (Ph.D. thesis). (April), The University of Sydney, Sydney, URL: https://ses.library.usyd.edu.au/handle/2123/576.

Best, V., Mason, C.R., Swaminathan, J., Roverud, E., Kidd, G., 2017. Use of a glimpsing model to understand the performance of listeners with and without hearing loss in spatialized speech mixtures. J. Acoust. Soc. Am. 141, 81–91. http://dx.doi.org/10.1121/1.4973620.

Bolia, R.S., Ericson, M.A., Nelson, W.T., McKinley, R.L., Simpson, B.D., 1999. A cocktail party effect in the median plane? J. Acoust. Soc. Am. 105 (2), 1390–1391. http://dx.doi.org/10.1121/1.426572.

Bolia, R.S., Nelson, W.T., Ericson, M.A., Simpson, B.D., 2000. A speech corpus for multitalker communications research. J. Acoust. Soc. Am. 107 (2), 1065–1066. http://dx.doi.org/10.1121/1.428288.

Borghini, G., Hazan, V., 2018. Listening effort during sentence processing is increased for non-native listeners: A pupillometry study. Front. Neurosci. Volume 12 - 2018, http://dx.doi.org/10.3389/fnins.2018.00152.

Bronkhorst, A.W., 2000. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. Acta Acust. United Acust. 86, 117–128.

Bronkhorst, A.W., 2015. The cocktail-party problem revisited: early processing and selection of multi-talker speech. Atten. Percept. Psychophys. 77, 1465–1487. http://dx.doi.org/10.3758/S13414-015-0882-9/FIGURES/1.

Brungart, D.S., 2001. Informational and energetic masking effects in the perception of two simultaneous talkers. J. Acoust. Soc. Am. 109, 2112. http://dx.doi.org/10.1121/1.1345696.

Cherry, E.C., 1953. Some experiments on the recognition of speech, with one and with 2 ears. J. Acoust. Soc. Am. 25, 975–979. http://dx.doi.org/10.1121/1.1907229.

Cuevas-Rodriguez, M., Gonzalez-Toledo, D., Reyes-Lecuona, A., Picinali, L., 2021. Impact of non-individualised head related transfer functions on speech-in-noise performances within a synthesised virtual environment. J. Acoust. Soc. Am. 149 (4), 2573–2586. http://dx.doi.org/10.1121/10.0004220.

Cuevas-Rodríguez, M., Picinali, L., González-Toledo, D., Garre, C., de la Rubia-Cuestas, E., Molina-Tanco, L., Reyes-Lecuona, A., 2019. 3D tune-in toolkit: An open-source library for real-time binaural spatialisation. In: Yasin, I. (Ed.), PLoS One 14 (3), e0211899. http://dx.doi.org/10.1371/journal.pone.0211899.

Culling, J.F., Hawley, M.L., Litovsky, R.Y., 2005. Erratum: The role head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources [J. Acoust. Soc. Am. 116, 1057 (2004)]. J. Acoust. Soc. Am. 118 (1), 552. http://dx.doi.org/10.1121/1.1925967.

Daugintis, R., Barumerli, R., Picinali, L., Geronazzo, M., 2023. Classifying non-individual head-related transfer functions with a computational auditory model: Calibration and metrics. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 1–5. http://dx.doi.org/10.1109/ICASSP49357.2023.10095152.

Durlach, N.I., 1972. Binaural signal detection: Equalization and cancellation theory. In: Tobias, J. (Ed.), In: Foundations of Modern Auditory Theory, vol. II, Academic, New York, pp. 371–462.

Edraki, A., Chan, W.-Y., Jensen, J., Fogerty, D., 2022. Spectro-temporal modulation glimpsing for speech intelligibility prediction. Hear. Res. 426, 108620. http://dx.doi.org/10.1016/j.heares.2022.108620.

Engel, I., Alon, D.L., Scheumann, K., Crukley, J., Mehra, R., 2022. On the differences in preferred headphone response for spatial and stereo content. J. Audio Eng. Soc. 70 (4), 271–283. http://dx.doi.org/10.17743/jaes.2022.0005.

Engel, I., Daugintis, R., Vicente, T., Hogg, A., Pauwels, J., Tournier, A., Picinali, L., 2023. The SONICOM HRTF dataset. J. Audio Eng. Soc. 71, 241–253. http://dx.doi.org/10.17743/jaes.2022.0066.

Freyman, R.L., Balakrishnan, U., Helfer, K.S., 2001. Spatial release from informational masking in speech recognition. J. Acoust. Soc. Am. 109 (5), 2112–2122. http://dx.doi.org/10.1121/1.1354984.

Freyman, R.L., Balakrishnan, U., Helfer, K.S., 2004. Effect of number of masking talkers and auditory priming on informational masking in speech recognition. J. Acoust. Soc. Am. 115, 2246–2256. http://dx.doi.org/10.1121/1.1689343.

Freyman, R.L., Helfer, K.S., McCall, D.D., Clifton, R.K., 1999. The role of perceived spatial separation in the unmasking of speech. J. Acoust. Soc. Am. 106, 3578–3588. http://dx.doi.org/10.1121/1.428211.

Garcia Lecumberri, M.L., Cooke, M., Cutler, A., 2010. Non-native speech perception in adverse conditions: A review. Speech Commun. 52, 864–886. http://dx.doi.org/10.1016/j.specom.2010.08.014.

González-Toledo, D., Cuevas-Rodríguez, M., Vicente, T., Picinali, L., Molina-Tanco, L., Reyes-Lecuona, A., 2024. Spatial release from masking in the median plane with non-native speakers using individual and mannequin head related transfer functions. J. Acoust. Soc. Am. 155 (1), 284–293. http://dx.doi.org/10.1121/10.0024239.

Hawley, M.L., Litovsky, R.Y., Culling, J.F., 2004. The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. J. Acoust. Soc. Am. 115, 833–843. http://dx.doi.org/10.1121/1.1639908.

Jelfs, S., Culling, J.F., Lavandier, M., 2011. Revision and validation of a binaural model for speech intelligibility in noise. Hear. Res. 275 (1), 96–104. http://dx.doi.org/10.1016/j.heares.2010.12.005.

Jenny, C., Reuter, C., 2020. Usability of individualized head-related transfer functions in virtual reality: Empirical study with perceptual attributes in sagittal plane sound localization. JMIR Serious Games 8 (3), e17576. http://dx.doi.org/10.2196/17576.

Kidd, J., Mason, C.R., Swaminathan, J., Roverud, E., Clayton, K.K., Best, V., 2016. Determining the energetic and informational components of speech-on-speech masking. J. Acoust. Soc. Am. 140 (1), 132–144. http://dx.doi.org/10.1121/1.4954748.

Kim, C., Lim, V., Picinali, L., 2020. Investigation into consistency of subjective and objective perceptual selection of non-individual head-related transfer functions. J. Audio Eng. Soc. 68, 819–831. http://dx.doi.org/10.17743/jaes.2020.0053.

Lavandier, M., Vicente, T., Prud'homme, L., 2022. A series of SNR-based speech intelligibility models in the auditory modeling toolbox. Acta Acust. 6, 20. http://dx.doi.org/10.1051/aacus/2022017.

Majdak, P., Hollomey, C., Baumgartner, R., 2022. AMT 1.x: A toolbox for reproducible research in auditory modeling. Acta Acust. 6, 19. http://dx.doi.org/10.1051/aacus/2022011.

Martin, R.L., McAnally, K.I., Bolia, R.S., Eberle, G., Brungart, D.S., 2012. Spatial release from speech-on-speech masking in the median sagittal plane. J. Acoust. Soc. Am. 131 (1), 378–385. http://dx.doi.org/10.1121/1.3669994.

McAnally, K.I., Bolia, R.S., Martin, R.L., Eberle, G., Brungart, D.S., 2002. Segregation of multiple talkers in the vertical plane: Implications for the design of a multiple talker display. In: Human Factors and Ergonomics Society Annual Meeting, vol. 46, no. 4, SAGE Publications, Los Angeles (CA), pp. 588–591. http://dx.doi.org/10.1177/154193120204600404.

Picinali, L., Katz, B.F.G., 2023. System-to-user and user-to-system adaptations in binaural audio. In: Geronazzo, M., Serafin, S. (Eds.), Sonic Interactions in Virtual Environments. Springer International Publishing, Cham, pp. 115–143. http://dx.doi.org/10.1007/978-3-031-04021-4{_}4.

Rosen, S., Souza, P., Ekelund, C., Majeed, A.A., 2013. Listening to speech in a background of other talkers: Effects of talker number and noise vocoding. J. Acoust. Soc. Am. 133 (4), 2431–2443. http://dx.doi.org/10.1121/1.4794379.

Scharenborg, O., van Os, M., 2019. Why listening in background noise is harder in a non-native language than in a native language: A review. Speech Commun. 108, 53–64. http://dx.doi.org/10.1016/J.SPECOM.2019.03.001.

Wasiuk, P.A., Buss, E., Oleson, J.J., Calandruccio, L., 2022. Predicting speech-in-speech recognition: Short-term audibility, talker sex, and listener factors. J. Acoust. Soc. Am. 152 (5), 3010–3024. http://dx.doi.org/10.1121/10.0015228.

Wasiuk, P.A., Calandruccio, L., Oleson, J.J., Buss, E., 2023. Predicting speech-in-speech recognition: Short-term audibility and spatial separation. J. Acoust. Soc. Am. 154 (3), 1827–1837. http://dx.doi.org/10.1121/10.0021069.